Research Paper | Computer Engineering, | E-ISSN No : 2454-9916 | Volume : 8 | Issue : 7 | Jul 2022

COVID-19 PREDICTION USING MACHINE LEARNING TECHNIQUES

Sayali R. Nipanikar[1], Dr. K. Rajeswari[1]

[1]Department of Computer Engineering, PCCOE, Pune, India.

ABSTRACT

The abundance of type and quantity of available data in the healthcare field has led many to utilize machine learning approaches to keep up with this influx of data. Data pertaining to COVID-19 is an area of recent interest. The widespread influence of the virus across the United States creates an obvious need to identify groups of individuals that are at an increased risk of mortality from the virus. We propose a so-called clustered random forest approach to predict COVID-19 patient mortality. We use this approach to examine the hidden heterogeneity of patient frailty by examining demographic information for COVID-19 patients. We find that our clustered random forest approach attains predictive performance comparable to other published methods. We also find that follow-up analysis with decision tree algorithms and linear regression provide insight into the type and magnitude of mortality risks associated with COVID-19.

KEYWORDS: Machine Learning, Random Forest, Decision Tree Algorithms.

I. INTRODUCTION:

Coronavirus (COVID-19) started in China in December 2019. As of January 2021, over 95 million cases have been reported around the world, with a mortality rate of 2% of the total closed cases [1]. This rapid pandemic expansion represents a global concern and a serious threat to the public health and economy worldwide. To prevent the infection from spreading, most countries restricted social interaction through precautionary measures such as isolation and quarantine. However, many infected patients did not benefit from the proper treatment due to late diagnosis and the novel and unknown nature of the virus. Recently, many researchers focused on developing new methodologies to screen infected patients in different stages to find notable associations between the patient's clinical features and the chances to succumb to the disease [2, 3]. Current investigation studies determined that artificial intelligence (AI) and machine learning (ML) techniques can play a key role in reducing the effect of the virus spread [4–6]. ML application technologies on patients' data fall under a range of different research directions [7]. One of the most important research directions is predicting the infection rate and mortality rate and building a model to classify patients based on their clinical findings [8, 9]. These research investigations are extremely important and would greatly assist people in the health sectors to be well prepared and take all necessary precautions to minimize the pandemic spread.

The aim of this research is to develop a prediction model to calculate the severity of the disease in COVID-19 patients, using risk factors that can be monitored remotely, with the patient being at home. Moreover, the study explores the impact of vital signs, chronic diseases, preliminary clinical investigations, and demographic features to predict the survival versus the mortality of COVID-19 patients. The study used COVID-19 patients' data from the King Fahad University Hospital containing the clinical findings and demographic information to validate the model perfor- mance and effectiveness. All the risk factors or vital signs that can be measured through widely used sensors were included in the study such as oxygen level in the blood, temperature, pulse rate, and blood pressure. The model will serve as an early warning system to timely identify at-risk patients.

II. RELATED WORK:

Early detection and diagnosis using AI techniques help to prevent the spread and to combat the COVID-19 pandemic using different data such as CT scans, X-ray, clinical data, and blood sample data.

Yan et al. [10] predicted the criticality and survival chances of patients with severe COVID-19 infection based on different risk factors and demographic information. The dataset used consists of 375 records from patients admitted to Tongji Hospital from January 10th to February 18th, 2020, including 201 survivors and 174 deceased within the same period. They used an XGBoost (XGB) model and identified only three main clinical features as significant, i.e., lactic dehydrogenase (LDH), lymphocyte, and high-sensitivity C-reactive protein (Hs-CRP), selected from more than 300 features. The proposed model was validated using data from 29 patients. The key findings of the research were the model's ability to predict the risk of death with 0.95 precision and 0.90 prediction accuracy. Such models will equip physicians with a tool for identifying critical conditions, thereby helping to reduce the mortality rate. Even though these findings are of great importance, the research has some limitations, which affect the accuracy of the reported results. These limitations were due to the small size of the dataset, namely, 29 records of patients only.

Similarly, Wong and So [11] also used XGB with another dataset to predict the severe and the death cases and identify the risk factors associated with COVID-19. The dataset was retrieved from United Kingdom Biobank (UKBB) and includes 93 different variables collected between 16 March 2020 and 19 July 2020. Two different studies have been conducted based on the sample's groups. For the first study, the data were clinical prediagnostic data of 1747 COVID-19 infected patient records containing both severe and death cases. For the severity class, the accuracy achieved was 0.668, and for the fatality class, the accuracy was 0.712. For the second study, the data were taken from the negative cases, the general population with no COVID-19 infection, con- sisting of 489987 records. The same model was applied, and the accuracy achieved was similar to the first study, with an accuracy of 0.669 for the severity class and 0.749 for the fatality class, respectively. It is worth mentioning that the researchers identified the five most significant risk factors for severe cases and death cases, with age being the top factor for both cases. Other factors include obesity, impaired renal function, multiple comorbidities, and cardiometabolic abnormalities.

III. DATA SET:

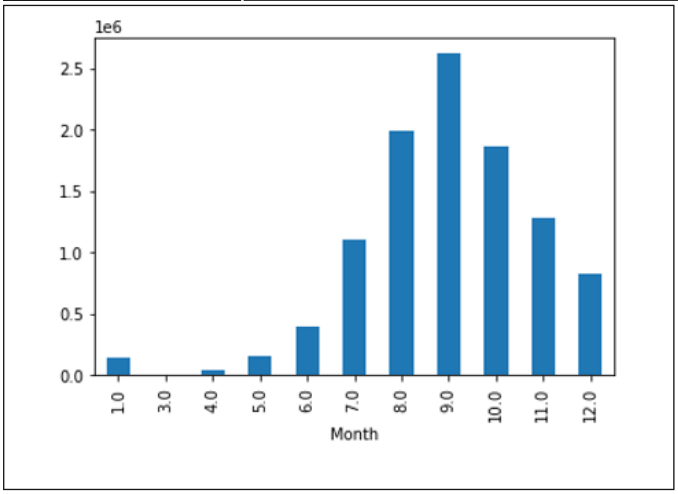
In this paper, I have used a dataset of more than 17365 laboratory-confirmed COVID-19 patients from 146 countries around the world including 307,382 labeled samples containing both male and female patients with an average age of 44.75. The disease was confirmed by detection of virus nucleic acid . The original dataset contained 32 data elements from each patient, including demographic and physiological data. At the data cleaning stage, we removed useless and redundant data elements such as data source, admin id, and admin name. We have also removed the unlabeled data samples. Then, data imputation techniques including mean/median/mode value replacement and KNN technique were used to handle missing values.

To have an accurate and unbiased model, we made sure that our dataset is balanced. A balanced dataset with an equal number of observations for both recovered and deceased patients was created to train and test our model. The data samples (patients) in the training dataset have been selected randomly and they are completely separate from the testing data.

IV. DATA UNDERSTANDING AND PRE-PROCESSING:

A. Data understanding:

The purpose is to create a model to estimate house prices. We split the set of knowledge into functions and target variables. In this section, we aim to understand the overview and original features of the original data set, and perform exploratory analysis of the information set to obtain useful observations [1]. This dataset contains quite a few categorical variables that need to be converted to numeric form using label encoding or creating dummy variables. These are real variable placeholders, fake/dummy variables you create yourself. Also, there are a lot of null values and outliers, so you need to handle them accordingly. Bath, prices, and balcony features are numeric variables. Represented by functional category variables such as area_type, total_sqft, location, society, availability, and size[1].



It can be seen that the price distribution is very different. The prices range from 8 lakhs to 3600 lakhs. Most values are less than 500 lakhs.

B. Data pre-processing:

Preprocessing is one of the key steps in data analysis and prediction. Several preprocessing tech- niques were applied on the dataset. The dataset contains data of all the patients admitted in the hospital. Some symptoms or vital signs occurred with very low frequency and were therefore removed from the dataset. All symptoms with occurrences at 50% or above were selected to be added to the feature set, while the symptoms with occurrences in the range from 2% to 49% were accumulated as one feature that was assigned a unique code. The first three vital signs: fever, cough, and shortness of breath (SOB) were defined as symptom features, while the remaining features were in- corporated as a new attribute "sym_others." 5% of the pa- tients in the study were asymptomatic at the time of initial diagnosis and considered as a part of the sym_others at- tribute. Similarly, the chronic top three (3) diseases (i.e., diabetes, high blood pressure, and cardiac) with the highest frequency were included as features. However, all other chronic disease types with more than 1 occurrence were incorporated as one feature "chr_others." After the initial preprocessing data, an encoding scheme was applied on the categorical features. As the dataset contains a small number of missing values, imputation was performed using the decision tree algorithm.

V . METHODOLOGY:

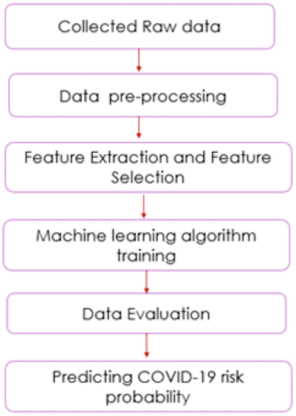


Fig. 2: Flowchart of Implementation

Based on Fig.2, the process of regression analysis and decision tree algorithm is described in the following section:

A. Basic Linear Regression Model:

Linear regression is based on supervised learning. It performs the tasks to predict a dependent variable value(Y) based on a given independent variable(X). It is the relationship between input (X) and output (Y). It is one of the most well- known and well-understood algorithms in machine learning[4]. A linear regression line has an equation of the form $Y = a + bX$, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when $x = 0$).

B. Random Forest Tree algorithm:

The Our method for predicting COVID-19 patient mortality in this project relies heavily on the random forest (RF) classifier from Breiman. Consequently, a brief description of this method is appropriate. The RF classifier is itself made up of many decision trees. A decision tree classifier is made by successively splitting our data at decision nodes according to feature values. Our initial decision node splits the data into two groups according to a cutoff value for one of the data features. Then these groups are again split by a decision node, and this process continues, building out the "branches" of the decisions tree. When the splitting stops, the last remaining groups, or "leaves" of the decision tree provide the designation for which class individuals in that group belong to. The feature used at each decision node to split the data is typically chosen so that error at that step is minimized. The RF classifier uses an ensemble "forest" of these decision trees to make its classifications. Each tree in the RF ensemble is built using a bootstrapped random sample of the available data and considering only a random selection of available features when each splitting node is made. To classify an observation with the RF model, each decision tree in the ensemble "votes" for the class it predicts and the majority vote of the decision trees in the ensemble is the class that the RF classifier predicts. This reliance on the majority vote to classify an observation provides for better performance than a single decision tree classifier.

C. Decision Tree Algorithm:

An object that trains a tree-structured model to predict future data in order to produce meaningful continuous output. Decision trees, steps related to regression are the basic concepts of decision trees, maximizing information acquisition, classification trees, and regression trees. The basic concept of a decision tree consists of recursive partitioning. The root node, known as the parent node, can split each node into child nodes. These nodes can be the parent node of the resulting child node. Optimization of information gain tree learning algorithms are defined as functional nodes useful for defining objective functions.

VI. CONCLUSION:

The system uses this data in the most efficient way. Linear regression algorithms help satisfy customers by increasing the accuracy of real estate selection and reducing the risk of real estate investments. Many features that can be added to make the system more widely accepted. One of the major future scopes is to add more city real estate databases. This allows users to explore more properties and make informed decisions. More factors should be added, such as a recession that affects house prices [2]. Add detailed details for all real estate to provide detailed information on the desired real estate sample. This helps the system to run at a higher level.

In this paper, an overview of the concept of machine learning along with its various applications is discussed[5]. Taking data samples for houses and considering its various attributes, house prices were predicted using machine learning regression methods to predict the price of the property by using previous data and to check quality of solution or output. Data modeling and analysis of these jobs have a range of for future applications in a flat value prediction system. Based on the results, it can be concluded that forecasts Machine Learning focused guess is understandably and meaningful to data analysis points from view. When done correctly the ratio can be achieved is high or is exactly, and thus Machine Learning techniques find applications in many fields.

VII. ACKNOWLEDGMENT:

It is a great honor to publish a paper on "COVID-19 Prediction Using Machine Learning Techniques". The project is the project guide Dr. Excellent guidance from K.Rajeswari. The project has received sincere support, inspiration, encouragement and valuable guidance at every step. Also with the help of Dr. N. B.Chopade, Principal of Pimpri chinchwad College of Engineering, was a huge help on this project. Finally, I would like to thank my classmates who provided ideas and support for the project, either directly or indirectly. We would also like to thank our pre-training staff for providing various facilities for the project.

REFERENCES:

I. Cao Y, Li L, Feng Z, Wan S, Huang P, Sun X, Wen F, Huang X, Ning G, Wang W. Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. Cell Discov 6: 11, 2020. doi:10.1038/s 41421-020-0147-1.

II. COVID-19. Open Research Dataset (CORD-19). 2020, https://pages. semanticscholar.org/coronavirus-research.

III. Fang L, Karakiulakis G, Roth M. Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? Lancet Respir Med In press, 2020. doi:10.1016/S2213-2600(20)30116-8.

IV. Ge Y, Tian T, Huang S, Wan F, Li J, Li S, Yang H, Hong L, Wu N, Yuan E, Cheng L, Lei Y, Shu H, Feng X, Jiang Z, Chi Y, Guo X, Cui L, Xiao L, Li Z, Yang C, Miao Z, Tang H, Chen L, Zeng H, Zhao D, Zhu F, Shen X, Zeng J. A data-driven drug repositioning framework discovered a potential therapeutic agent targeting COVID-19. bioRxiv, 2020. doi:10.1101/2020.03.11.986836.

V. Gozes O, Frid-Adar M, Greenspan H, Browning PD, Zhang H, Ji W, Bernheim A, Siegel E. Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection & Patient Mon- itoring using Deep Learning CT Image Analysis. arXiv2003.05037. 2020.

VI. Metsky HC, Freije CA, Kosoko-Thoroddsen T-SF, Sabeti PC, Myhr- vold C. CRISPR-based COVID-19 surveillance using a genomically- comprehensive machine learning approach. bioRxiv, 2020. doi:10.1101/ 2020.02.26.967026.

VII. Ong E, Wong MU, Huffman A, He Y. COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. bioRxiv. 2020. doi:10.1101/2020. 03.20.000141.

VIII. Randhawa GS, Soltysiak MPM, El Roz H, de Souza CPE, Hill KA, Kari L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. bioRxiv, 2020.

IX. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D. Improved protein structure prediction using potentials from deep learning. Nature 577: 706 –710, 2020. doi:10.1038/s41586-019- 1923-7.

X. Wang Y, Hu M, Li Q, Zhang X-P, Zhai G, Yao N. Abnormal respiratory patterns classifier may contribute to large-scale screening of people in- fected with COVID-19 in an accurate and unobtrusive manner. arXiv2002.05534. 2020.

XI. Yan L, Zhang H-T, Xiao Y, Wang M, Sun C, Liang J, Li S, Zhang M, Guo Y, Xiao Y. Prediction of survival for severe Covid-19 patients with three clinical features: development of a machine learning-based prognos- tic model with clinical data in Wuhan. medRxiv. 2020. doi:10.1101/2020. 02.27.20028027.

XII. Zhavoronkov A, Aladinskiy V, Zhebrak A, Zagribelnyy B, Terentiev V, Bezrukov DS, Polykovskiy D, Shayakhmetov R, Filimonov A, Orekhov P. Potential COVID-2019 3C-like Protease Inhibitors Designed Using Generative Deep Learning Approaches. Insilico Med Hong Kong Ltd A 307: E1, 2020.

XIII. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-L, Chen HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579: 270 –273, 2020. doi:10.1038/s41586-020-2012-7.

XIV. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0252384

XV. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8138040/

XVI. https://www.medrxiv.org/content/10.1101/2021.01.29.21250762v1.full.pdf

XVII. (engpaper.com)

XVIII. COVID-19 Future Forecasting Using Supervised Machine Learning Models | IEEE Journals & Magazine | IEEE Xplore

XIX. Using Automated Machine Learning to Predict the Mortality of Patients With COVID-19: Prediction Model Development Study (nih.gov)

XX. Prediction of COVID-19 Using Genetic Deep Learning Convolutional Neural Network (GDCNN) | IEEE Journals & Magazine | IEEE Xplore